# 미디어에 나타난 부산 교통 관련 빅데이터의 분석

반재훈, 김용수, 이예찬, 정윤성, 정동민, 조해찬

고신대학교 IT경영학과

# Analysis of Transportation Big Data in Busan on Media

ChaeHoon Ban · YongSu Kim, YeChan Lee, YoonSeung Jung, DongMin Jeong, HaeChan Cho

Dept. of IT Management, Kosin University

E-mail : chban@kosin.ac.kr · ehyerisin@gmail.com · ycl0188@naver.com · vision8510@naver.com · jdmin96@naver.com · chc0451@naver.com

요 약

정보기술과 디지털 경제의 확산으로 대규모의 데이터가 생산되는 정보화시대에서 빅데이터의 중요성이 강조되고 있으며 다양한 분야에서 이를 응용하고 있다. 빅 데이터 분석 도구인 R은 통계 기반의 정보 분석을 가능하게 하는 언어와 환경이다. 본 논문에서는 R을 이용하여 미디어에 나타난 부산 교통 관련 빅데이터를 분석한다. 다양한 미디어에서 부산 교통 관련 데이터를 수집하고 어떠한 텍스트가 분포되어 있는지 빈도 조사를 수행한다.

키워드

Big Data, R, Text Mining, Transportation, Analysis

## Ⅰ. 서론

정보기술과 디지털 경제의 확산으로 대규모의 데이터가 생산되는 정보화시대에서 빅데이터의 중요성이 강조되고 있으며 다양한 분야에서 이를 응용하고 있다. 빅 데이터 분석 도구인 R은 통계 기반의 정보 분석을 가능하게 하는 언어와 환경이다. 최근 무인자동차 산업의 기술이 큰 이슈가 되고 있다. 무인자동차는 교통정보, 지도, 도로현황 등 교통관련 빅데이터가 핵심적으로 포함되어 있어 교통관련 빅데이터의 연구를 중심으로 미래 교통산업 자원으로 활용됨으로 빅데이터에 의미가 중요하다고 볼 수 있다.

부산시와 교통과 관련된 키워드를 중심으로 최근 1년간, 5년간 총 4개의 신문사(경남신문, 국제신문, 부산일보, 중앙일보)의 기사를 각 신문사 분석과 4개의 신문사를 종합한 기사의 데이터를 분석하고 비교하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 부산교통관련 빅데이터 기법에 관련된 연구를 기술한다. 3장에서는 본 논문에서 구현한 워드 클라우드 형태의 그림을 표현하기 위해 R 프로그램을 활용한 데이터 분석 방법에 대해 기술한다. 4장에서는 워드 클라우드 형태의 그림으로 표현한 각 신문사의 결과와 비교를 설명하고, 마지막 5장에서는 결론 및 향후 연구에 대해 기술한다.

## Ⅱ. 관련연구

기존의 연구에서는 데이터 마이닝, 텍스트 마이닝, 오피니언 마이닝, 웹 마이닝, 소셜 마이닝 기법 등 다양한 기법을 통한 빅 데이터 분석연구가 있었다. 정보통신의 발달과 소셜 미디어의 급속한 확산으로 빅 데이터가 경제적으로 자산이 되고 있는 시대를 맞이하는 데 필요한 데이터 분석기법과 인프라 기술에 대해 알아보고, 한글 Text 데이터를 R 프로그램을 이용하여 usesejongdic() 이라는 옵션을 이용하여 명사만 추출하는 방법으로 비정형 데이터를 분석하였다.[1] 데이터 시각화 도구 통계 패키지인 R을 이용하여 대기오염의 자료를 여러 가지 방법의 데이터 시각화를 통하여 나타내었고, 데이터 시각화 방법별로 통계적인 방법을 활용한 분석과 연계하여 어떤 특징이 있는지를 나타냈다. 2차원의 히스토그램과 선점도, 상자그림, 3차원 산점도와 투시도 등 다양한 방법의 그래프를 구현하여 오존농도와 설명 변수들 간에 어떠한 관련성이 있는지를 분석했다.[2] 빅데이터 분석 도구인 R을 이용하여 빠른 시간 안에 사용자가 목적으로 하고 있는 특허검색 결과를 효율적으로 도출할 수 있는 검색어 추출에 관한 연구를 진행했다.[3] 데이터 마이닝의 일부인 텍스트 마이닝의 기법을 이용하여 부산지역지인 국제신문과 부산일보의 기사들 중 제목에 ‘부산’과 ‘교통’을 동사에 포함한 기사의

기사 내용의 관계 또는 관련 있는 데이터에 내재되어 있는 의미 있는 패턴을 찾는 사회네트워크 분석을 실시하여 정형화된 빅 데이터를 시각화하고 해석했다.[4] 구글, 야후, 네이버 등 주요 포털의 지도에는 POI(Point of interest)가 서비스 되고 있다. 지도의 위치 데이터 즉, 현재 이용자가 위치한 장소는 인문학적인 스토리텔링의 시작점을 주목하여, POI는 카페, 레스토랑, 병원, 식당 등의 정보만이 서비스되는 한계점을 지적하고, 더 나아가 대안으로 POI 정보와 결합된 소위 '인문융합 지도 서비스'를 제안 했다.[5] 빅데이터 분석도구인 R을 이용하여 성경의 텍스트 데이터를 성경전체, 구약성경, 신약성경, 모세오경, 사복음서 데이터 분석결과를 각각의 워드 클라우드 형태 그림으로 표현하여 성경데이터를 분석하여 성경을 읽는 독자에게 주는 메시지가 무엇인지에 대한 연구를 제시하였다.

## III. 데이터 분석 방법

빅데이터 분석도구인 R을 이용하여 텍스트 데이터를 워드 클라우드 형태의 그림으로 표현한다. 신문기사의 데이터는 검색 포털 사이트를 이용하여 '부산','교통'관련 키워드를 검색하여 본문내용을 중심으로 스크랩하여 텍스트 파일의 데이터를 수집햇으며, 데이터를 분석하기 위해 경남신문, 국제신문, 부산일보, 중앙일보를 최근 1년간, 5년간 구분하였으며 마지막으로 4개의 신문사의 데이터를 통합하여 분석하였다. 데이터의 분석과정은 그림1과 같다.

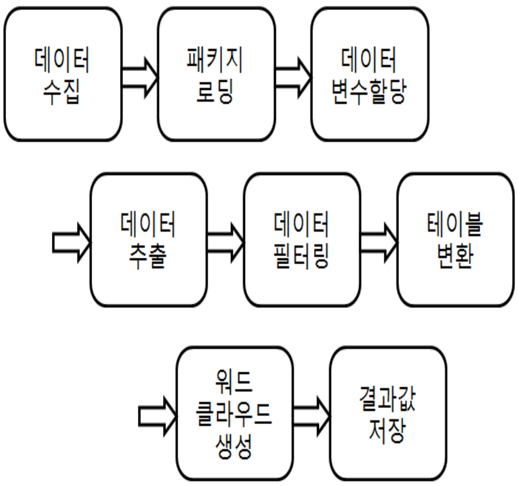


그림 1. 데이터 분석 과정

데이터 분석도구인 R을 설치하고 한글 데이터 분석에 필요한 패키지("KoNLP"), 워드 클라우드 생성에 필요한 패키지("wordcloud")를 설치하고 R 소스에 로딩한다. 수집한 데이터를 경남신문, 국제신문, 부산일보, 중앙일보, 전체기사의 그룹으로 분류하여 각 그룹의 데이터를 변수를 할당하여 대입한다. 한글의 명사를 추출해주는 함수인 'extracNoun'함수를 사용함으로서 성경 데이터를 명사로 변환하여 변환된 데이터를 확인 후 원하지 않는 데이터에 대한 'gsub'함수를 이용하여 데이터를 필터링 한다. 여기서는 2자리 이상의 명사만 추출하도록 프로그램을 구현하였다. 필터링 된 데이터를 텍스트 형식의 파일로 저장하여 테이블 형태로 변환하여 변수에 할당한다. 텍스트 형태로 각 명사에 대한 빈도수를 측정하여, 상위30위의 결과를 워드 클라우드 형태의 그래픽으로 출력한다. 출력 결과물을 이미지파일(JPGE, BMP, PNG 등)으로 저장한다.

## IV. 데이터 분석 결과 및 비교

본 논문에서는 '부산', '교통' 관련 데이터 분석의 결과를 워드 클라우드와 키워드 빈도 수에 대하여 표현하였다. 워드 클라우드란 문서의 키워드, 개념 등을 직관적으로 파악할 수 있도록 핵심단어를 시각적으로 돋보이게 하는 기법이다. 예를 들면 텍스트가 많이 언급될수록 단어를 크게 표현해 한눈에 들어올 수 있게 하는 기법 등이 있다.


그림 2. 경남신문 최근1년(좌), 최근5년(우)


그림 3. 국제신문 최근1년간(좌), 최근5년간(우)

그림2, 그림3, 그림4, 그림5와 같이 각각의 신문사에서 최근 1년간과 5년간의 상위 30단어를 데이터를 추출하였다. 그림6은 경남일보, 국제신문, 부산일보, 중앙일보의 모든 데이터를 통합하여 상위30단어를 추출하였다.


그림 4. 부산일보 최근1년간(좌), 최근5년간(우)


그림 5. 중앙일보 최근1년간(좌), 최근5년간(우)


그림 6. 신문종합 최근1년간(좌), 최근5년간(우)

표1은 4개의 신문사의 최근 1년간 상위 30단어의 키워드 빈도를 표시하였고, 표2는 최근 5년간의 상위 30단의 키워드 빈도를 표시하였다.

표 1. 신문종합 최근1년간 키워드 빈도

| | | | | |
|---|---|---|---|---|
| 부산 | 시간 | 구간 | 고속도로 | 교통 |
| 661 | 388 | 296 | 295 | 291 |
| 차량 | 버스 | 도시 | 지역 | 이용 |
| 262 | 248 | 243 | 240 | 225 |
| 사업 | 노선 | 부산시 | 국토 | 대중교통 |
| 224 | 218 | 209 | 191 | 191 |
| 운행 | 지난해 | 중앙 | 정체 | 예상 |
| 175 | 146 | 145 | 141 | 138 |
| 시민 | 오후 | 철도 | 안전 | 설치 |
| 134 | 133 | 129 | 128 | 124 |
| 서울 | 인상 | 공사 | 기준 | 정부 |
| 123 | 122 | 121 | 121 | 121 |

표 2. 신문종합 최근5년간 키워드 빈도

| | | | | |
|---|---|---|---|---|
| 부산 | 버스 | 도시 | 교통 | 시간 |
| 1318 | 519 | 480 | 466 | 463 |
| 차량 | 부산시 | 구간 | 사업 | 대중교통 |
| 444 | 419 | 406 | 352 | 335 |
| 철도 | 이용 | 운행 | 고속도로 | 지역 |
| 335 | 315 | 309 | 297 | 293 |
| 시민 | 노선 | 오후 | 개통 | 운영 |
| 289 | 261 | 235 | 221 | 215 |
| 정체 | 추진 | 안전 | 공사 | 지난해 |
| 199 | 198 | 195 | 194 | 190 |
| 교통공사 | 예상 | 오전 | 시내 | 국토 |
| 189 | 189 | 189 | 188 | 174 |

표2에서 고속도로의 빈도는 297회, 표1에서 고속도로의 빈도는 295회로 두 빈도수는 2회 차이를 보임으로 최근 고속도로에 관한 기사들과 고속도로에 관한 문제가 발생하였음을 엿볼 수 있었다. 표2에서는 주로 대중교통에 관한 단어들이 주로 나타났으며, 표1에서는 대중교통도 나타난 반면 '고속도로', '구간', '시간' 등의 단어가 언급되면서 부산시의 교통이 대중교통과 자가 소유의 자동차를 이용이 증가하였거나 자동차에 관한 문제들을 실었다는 것을 예측할 수 있다. 또한 4개의 신문사에서는 빈도수는 낮으나 '지난해'의 단어를 언급함으로 주로 전년도와 비교하는 기사들이 많았다는 것을 알 수 있다.

## Ⅴ. 결론 및 향후 연구

본 논문에서는 정보기술과 디지털 경제의 확산으로 대규모의 데이터가 생산되는 정보화시대에서 빅데이터의 중요성이 강조되고 있으며 다양한 분야에서 응용하고 있다. 빅데이터 분석도구 R을 이용하여 미디어 매체에 나타난 부산 교통관련 빅데이터를 워드클라우드 형태의 그림으로 나타내어 경남신문, 국제신문, 부산일보, 중앙일보에서 나타는 키워드를 워드 클라우드로 시각화함으로써 빈도 수에 따른 키워드를 쉽게 알아 볼 수 있었다. R프로그램을 이용함으로써 누구나 쉽게 접근하여 다양한 데이터를 워드 클라우드 형태의 그림으로 시각화 표현을 구현할 수 있다고 본다.

향후 연구 방향으로서 더욱 많은 빅데이터, 부산 교통관련 데이터를 수집하여 앞으로 무인자동차시스템의 기술에 각 도시의 교통상황을 접목하여 무인자동차의 운행 기술에 도움이 될 수 있으며, 앞으로의 문제점과 해결방안에 대한 예측을 시도할 수 있고, 다양한 분야에 응용할 수 있을 것이다.

## 참고문헌

[1] 김현근. R을 이용한 빅 데이터 사례 분석. 호서대학교 일반대학원 정보통계학과 석사학위 논문, 2014.

[2] 오영창, 박은식. 2015. R 소프트웨어를 이용한 대기오염 데이터의 시각화. 한국데이터정보과학회지, 26(2), 399-408

[3] 장청윤, 장정환, 김석주, 이현균,& 이창호. (2013). 빅데이터 분석 도구 R을 활용한 효율적인 특허 검색에 관한 연구. 대한안전경영과학회지, 15(4), 289-294.

[4] 이경준, 노윤환, 윤상경, 조영석. 2014. 부산지역 교통관련 기사를 이용한 비정형 빅데이터의 정형화와 시각적 해석. 한국데이터정보과학회지, 25(6), 1431-1438

[5] 이원태, 강장묵. 2015. 빅데이터 중 POI와 공간 메타포를 활용한 인문 융합 지도 연구. 한국인터넷방송통신학회. 15(3), 43-50

[6] 김용수, 반재훈 2016. 성경 데이터를 활용한 빅데이터 분석. 한국정보통신학회 2015 추계 종합학술대회, 349-352